1. What does it mean for a random variable $W \in \mathbb{R}$ to be sub-Gaussian with parameter $\sigma > 0$? State an upper bound on $\mathbb{P}(W - \mathbb{E}W > t)$ for $t > 0$.

   Show that if $W_1, \ldots, W_n$ are independent and sub-Gaussian with parameter σ, then $\sum_{i=1}^n W_i/n$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$.

   State Hoeffding's Lemma.

   Now suppose matrix $X \in [-1,1]^{n\times p}$ with $p \geq 2$ has independent rows with $\mathbb{E}(X_{ij}) = 0$ and $\mathbb{E}(X_{ij}X_{ik}) = \Sigma_{jk}$ for all i, j, k and positive definite matrix $\Sigma \in \mathbb{R}^{p\times p}$. Let $\hat{\Sigma} = X^TX/n$. Show that with probability at least $1 - 2p^{-2}$,

$$\max_{j,k} |\hat{\Sigma}_{jk} - \Sigma_{jk}| \leq 2\sqrt{2\log(p)/n}.$$

   Solution: A random variable W is *sub-Gaussian* with parameter $\sigma > 0$ if

$$\mathbb{E}e^{\alpha(W-\mathbb{E}W)} \leq e^{\alpha^2\sigma^2/2} \quad \text{for all } \alpha \in \mathbb{R}.$$

   Have $\mathbb{P}(W - \mathbb{E}W > t) \leq e^{-t^2/(2\sigma^2)}$. Suppose wlog $\mathbb{E}W_i = 0$ for each i. Then

$$\begin{aligned}\mathbb{E}\exp\Big(\alpha\sum_{i=1}^n W_i/n\Big) &= \prod_{i=1}^n \mathbb{E}\exp(\alpha W_i/n)\\ &\leq \prod_{i=1}^n \exp(\alpha^2\sigma^2/(2n^2))\\ &= \exp\Big(\alpha^2\sigma^2/(2n)\Big),\end{aligned}$$

   showing $\sum_{i=1}^n W_i/n$ is sub-Gaussian with parameter $\sigma/\sqrt{n}$.

   Hoeffding's lemma: if W takes values in $[a, b]$ then W is sub-Gaussian with parameter $(b - a)/2$.

   Each $X_{ij}X_{ik} \in [-1, 1]$, so from the above, $\hat{\Sigma}_{jk}$ is sub-Gaussian with parameter $1/\sqrt{n}$. Thus

$$\mathbb{P}(\hat{\Sigma}_{jk} - \Sigma_{jk} > 2\sqrt{2\log(p)/n}) \leq p^{-4}.$$

   Similarly, $\mathbb{P}(\Sigma_{jk} - \hat{\Sigma}_{jk} > 2\sqrt{2\log(p)/n}) \leq p^{-4}$, so $\mathbb{P}(|\hat{\Sigma}_{jk} - \Sigma_{jk}| > 2\sqrt{2\log(p)/n}) \leq 2p^{-4}$. Then

$$\begin{aligned}\mathbb{P}(\max_{jk} |\hat{\Sigma}_{jk} - \Sigma_{jk}| > 2\sqrt{2\log(p)/n}) &= \mathbb{P}(\cup_{jk}|\hat{\Sigma}_{jk} - \Sigma_{jk}| > 2\sqrt{2\log(p)/n})\\ &\leq 2p^{-2}.\end{aligned}$$

2. Suppose we have input–output pairs $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{-1, 1\}$. Consider the empirical risk minimisation problem using hinge loss and hypothesis class

$$\mathcal{H} = \{x \mapsto x^T\beta : \beta \in C \subseteq \mathbb{R}^p\},$$

   where C is a non-empty closed convex set. Briefly explain why the objective function $f : \mathbb{R}^p \to \mathbb{R}$ of the optimisation problem is convex.

Now take $C = \{x \in \mathbb{R}^p : x_j \geq 0 \text{ for } j = 1, \ldots, p\}$. Write down the (sub)gradient descent procedure for minimising f over $\beta \in C$ giving explicit forms for any subgradients and projections used.

Let $\hat{\beta} \in \mathbb{R}^p$ be a minimiser f over C and suppose that $\max_{i=1,\ldots,n} \|x_i\|_2 \leq M$. Prove that the output $\bar{\beta}$ of your procedure with k iterations initialised at a $\beta_1 \in \mathbb{R}^p$ and implemented with a fixed step size η you should specify satisfies

$$f(\bar{\beta}) - f(\hat{\beta}) \leq \frac{M\|\hat{\beta} - \beta_1\|_2}{\sqrt{k}}.$$

Solution: We have

$$f(\beta) = \frac{1}{n}\sum_{i=1}^{n} \max(1 - y_i x_i^T \beta, 0)$$

$\beta \mapsto \max(1 - y_i x_i^T \beta, 0)$ is convex as a maximum of linear (and hence convex) functions. Sums of convex functions with non-negative coefficients are convex, so f is convex.

Projection $\pi_C(x)$ minimises $\|z - x\|_2^2$ subject to $z_j \geq 0$ over $z \in \mathbb{R}^p$, so z_j minimises $(x_j - z_j)^2$ s.t. $z_j \geq 0$; thus $\pi_C(x)_j = \max(x_j, 0)$.

Subgradients of $\beta \mapsto \max(1 - y_i x_i^T \beta, 0)$ are $-y_i x_i t$ where $t = 1, 0$ if $y_i x_i^T \beta < 1$, $y_i x_i^T \beta > 1$ respectively, and $t \in [0, 1]$ if $y_i x_i^T \beta = 1$. Hence elements of $\partial f(\beta)$ take the form

$$-\frac{1}{n}\sum_{i=1}^{n} y_i x_i t_i$$

where

$$t_i \in \begin{cases} \{0\} & \text{if } y_i x_i^T \beta > 1 \\ [0, 1] & \text{if } y_i x_i^T \beta = 1 \\ \{1\} & \text{if } y_i x_i^T \beta < 1. \end{cases}$$

Gradient descent: Input any $\beta_1 \in \mathbb{R}^p$ such that $\beta \in C$, no. iterations k and positive step sizes $(\eta_s)_{s=1}^k$. For $s = 1$ to $k - 1$ do:
Take subgradient $g_s \in \partial f(\beta_s)$ of the form above
$z_{s+1} = \beta_s - \eta_s g_s$
$\beta_{s+1} = \pi_C(z_{s+1})$
Output $\bar{\beta} = \frac{1}{k}\sum_{s=1}^{k} \beta_s$.

Proof of convergence (η will be specified at the end): We have

$$\begin{aligned} f(\beta_s) - f(\hat{\beta}) &\leq g_s^T(\beta_s - \hat{\beta}) \\ &= -\frac{1}{\eta}(z_{s+1} - \beta_s)^T(\beta_s - \hat{\beta}) \\ &= \frac{1}{2\eta}\{\|\beta_s - z_{s+1}\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|z_{s+1} - \hat{\beta}\|_2^2\}. \end{aligned} \tag{1}$$

Projections are non-expansive: $\|\pi_C(z) - \pi_C(x)\|_2 \leq \|z - x\|_2$. Thus

$$\|z_{s+1} - \hat{\beta}\|_2^2 \geq \|\beta_{s+1} - \hat{\beta}\|_2^2.$$

Using this and (1),

$$f(\beta_s) - f(\hat{\beta}) \leq \frac{1}{2\eta}\{\eta^2\|g_s\|_2^2 + \|\beta_s - \hat{\beta}\|_2^2 - \|\beta_{s+1} - \hat{\beta}\|_2^2\}. \tag{2}$$

Now as $\|x_i\|_2 \leq M$, $\|g_s\|_2 \leq M$. Thus summing we get

$$\frac{1}{k}\sum_{s=1}^{k} f(\beta_s) - f(\hat{\beta}) \leq \frac{\eta M^2}{2} + \frac{1}{2\eta k}\left(\|\beta_1 - \hat{\beta}\|_2^2\right).$$

Taking the minimising $\eta = \|\beta_1 - \hat{\beta}\|_2/(M\sqrt{k})$ and using Jensen's inequality to give $f(\bar{\beta}) \leq \frac{1}{k}\sum_{s=1}^{k} f(\beta_s)$, we get the result.

3. Given a hypothesis class $\mathcal{H}$ of function $h : \mathcal{X} \to \mathbb{R}$ and i.i.d. input–output pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathcal{X} \times \{-1, 1\}$, define the *Rademacher complexity* $\mathcal{R}_n(\mathcal{H})$.

   Now suppose

$$\mathcal{H} = \Big\{x \mapsto \sum_{j=1}^{d} \beta_j \phi_j(x) : \beta \in \mathbb{R}^d \text{ and } \sum_{j=1}^{d} \gamma_j^2 \beta_j^2 \leq \lambda^2\Big\}.$$

   where $\phi_j : \mathcal{X} \to \mathbb{R}$ and $\gamma_j > 0$ for $j = 1, \ldots, d$. Let $C^2 = \mathbb{E}\left(\sum_{j=1}^{d}\{\phi_j(X_1)/\gamma_j\}^2\right)$. Show that

$$\mathcal{R}_n(\mathcal{H}) \leq \frac{\lambda C}{\sqrt{n}}.$$

   Let R_ϕ and $\hat{R}_\phi$ be the risk and empirical risk respectively for logistic loss, and h^* and $\hat{h}$ be the respective minimisers over $\mathcal{H}$ (so $\hat{h}$ is the empirical risk minimiser). Show that

$$\mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) \leq \frac{2\lambda C}{\log(2)\sqrt{n}}.$$

Solution:

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(X_i)\right)$$

where $\varepsilon_1, \ldots, \varepsilon_n$ is an i.i.d. Rademacher sequence independent of $X_{1:n}$.

Let $\mathcal{B}$ be the constraint set for β in the definition of $\mathcal{H}$. We have

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H}) &= \frac{1}{n}\mathbb{E}\left(\sup_{\beta\in\mathcal{B}} \sum_{i=1}^{n}\sum_{j=1}^{d} \beta_j \varepsilon_i \phi_j(X_i)\right) \\
&= \frac{1}{n}\mathbb{E}\left(\sup_{\beta\in\mathcal{B}} \sum_{j=1}^{d} \gamma_j\beta_j \sum_{i=1}^{n} \frac{\varepsilon_i \phi_j(X_i)}{\gamma_j}\right) \\
&\leq \frac{\lambda}{n}\mathbb{E}\left(\sum_{j=1}^{d}\left(\sum_{i=1}^{n} \frac{\varepsilon_i\phi_j(X_i)}{\gamma_j}\right)^2\right)^{1/2} \qquad \text{Cauchy–Schwarz} \\
&\leq \frac{\lambda}{n}\left(\sum_{j=1}^{d}\mathbb{E}\left(\sum_{i=1}^{n} \frac{\varepsilon_i\phi_j(X_i)}{\gamma_j}\right)^2\right)^{1/2} \qquad \text{Jensen applied to } \sqrt{\cdot}.
\end{aligned}$$

Now for $i \neq k$, $\mathbb{E}\{\varepsilon_i\varepsilon_k\phi_j(X_i)\phi_j(X_k)\} = 0$, so the above display is

$$\frac{\lambda}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{d}\mathbb{E}\{\phi_j(X_i)/\gamma_j\}^2\right)^{1/2} = \frac{\lambda C}{\sqrt{n}}.$$

Result: Contraction lemma if $\mathcal{F} := \{(x, y) \mapsto \phi(yh(x)) : h \in \mathcal{H}\}$ and $|\phi(u) - \phi(u')| \leq L|u - u'|$ then $\mathcal{R}_n(\mathcal{F}) \leq L\mathcal{R}_n(\mathcal{H})$.

Have $|\phi'(u)| \leq 1/\log 2$ for all u $\mathcal{R}_n(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{H})/\log(2)$. Result: For any class $\mathcal{F}$ of functions $f : \mathcal{Z} \to \mathbb{R}$ and i.i.d. random elements $Z_1, \ldots, Z_n$ taking values in $\mathcal{Z}$,

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \{f(Z_i) - \mathbb{E}f(Z_i)\} \right) \leq 2\mathcal{R}_n(\mathcal{F}).$$

Applying this to $-\mathcal{F} := \{-f : f \in \mathcal{F}\}$ for our $\mathcal{F}$ and noting that $\mathcal{R}_n(\mathcal{F}) = \mathcal{R}_n(-\mathcal{F})$, we get that

$$\mathbb{E} \sup_{h \in \mathcal{H}} \{R_\phi(h) - \hat{R}_\phi(h)\} \leq \frac{2\lambda C}{\log(2)\sqrt{n}}.$$

But then

$$\begin{aligned} \mathbb{E}R_\phi(\hat{h}) - R_\phi(h^*) &= \mathbb{E}\{R_\phi(\hat{h}) - \hat{R}_\phi(\hat{h})\} + \mathbb{E}\{\hat{R}_\phi(\hat{h}) - \hat{R}_\phi(h^*)\} + \mathbb{E}\{\hat{R}_\phi(h^*) - R_\phi(h^*)\} \\ &\leq \mathbb{E} \sup_{h \in \mathcal{H}} \{R_\phi(h) - \hat{R}_\phi(h)\} \leq \frac{2\lambda C}{\log(2)\sqrt{n}}. \end{aligned}$$

4. Let $\mathcal{F}$ be a family of functions $f : \mathcal{Z} \to \{a, b\}$ with $a \neq b$. Given $z_{1:n} \in \mathcal{Z}^n$, let $\mathcal{F}(z_{1:n}) = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$. What is the *empirical Rademacher complexity* $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ of $\mathcal{H}$? What is meant by the *VC dimension* $\mathrm{VC}(\mathcal{F})$ of $\mathcal{F}$?

   Now suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^p \times \{-1, 1\}$ are i.i.d. input–output pairs and consider performing empirical risk minimisation with misclassification loss over a class of classifiers $\mathcal{H}$. Let R and $\hat{R}$ denote the risk and empirical risk respectively. State an upper bound of $\mathbb{E}\sup_{h \in \mathcal{H}}(R(h) - \hat{R}(h))$ in terms of the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ of a class $\mathcal{F}$ related to $\mathcal{H}$ in a way you should specify.

   Let $\mathcal{B}$ be a family of functions $\phi : \mathbb{R} \to \{-1, 1\}$ given by

$$\mathcal{B} = \{u \mapsto \mathrm{sgn}(u - a),\ u \mapsto \mathrm{sgn}(a - u) : a \in \mathbb{R}\}.$$

   Compute $\mathrm{VC}(\mathcal{B})$. Let $u_1, \ldots, u_n \in \mathbb{R}$ and state an upper bound on $|\mathcal{B}(u_{1:n})|$.

   Now for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p) \in \mathcal{B}^p$ define $\mathcal{H}_{\boldsymbol{\phi}}$ by

$$\mathcal{H}_{\boldsymbol{\phi}} = \{v \mapsto \mathrm{sgn}(\beta_1\phi_1(v_1) + \cdots + \beta_p\phi_p(v_p)) : \beta_1, \ldots, \beta_p \in \mathbb{R}\}.$$

   Fix $x_1, \ldots, x_n \in \mathbb{R}^p$, and derive an upper bound on $|\mathcal{H}_{\boldsymbol{\phi}}(x_{1:n})|$.

   Let $\mathcal{H} := \cup_{\boldsymbol{\phi} \in \mathcal{B}^p} \mathcal{H}_{\boldsymbol{\phi}}$ and show that

$$|\mathcal{H}(x_{1:n})| \leq (n + 1)^{3p}.$$

   Finally conclude that

$$\mathbb{E} \sup_{h \in \mathcal{H}} (R(h) - \hat{R}(h)) \leq 2\sqrt{\frac{6p \log(n + 1)}{n}}.$$

Solution: Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. Rademacher random variables.

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \mathbb{E}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(z_i) \right).$$

$$\mathrm{VC}(\mathcal{F}) = \sup\{n \in \mathbb{N} : \max_{z_{1:n} \in \mathcal{Z}^n} |\mathcal{F}(z_{1:n})|\}$$

We have

$$\mathbb{E} \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} \leq 2\mathcal{R}_n(\mathcal{F}).$$

where $\mathcal{F} = \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ and ℓ denotes misclassification loss.

Clearly $\{0, 2\}$ can be shattered (e.g. $u \mapsto \mathrm{sgn}(u-1), \mathrm{sgn}(1-u), \mathrm{sgn}(u+1), \mathrm{sgn}(u-3)$ all map to each of the 2^2 subsets of $\{-1, 1\}$ when applied to $\{0, 2\}$). However for $u_{1:3} \in \mathbb{R}^3$ where wlog $u_1 \leq u_2 \leq u_3$, we can never have $f(u_1) = f(u_3) = 1 = -f(u_2)$ for $f \in \mathcal{B}$. Thus $\mathrm{VC}(\mathcal{B}) = 2$.

Sauer–Shelah: $|\mathcal{B}(u_{1:n})| \leq (n+1)^2$.

Result: For a vector space of functions $\mathcal{F}_1$, the class $\mathcal{G} = \{g : g(x) = \mathrm{sgn}(f(x)) : f \in \mathcal{F}_1\}$ has $\mathrm{VC}(\mathcal{G}) \leq \dim(\mathcal{F})$. Thus $\mathrm{VC}(\mathcal{H}_{\boldsymbol{\phi}}) \leq p$, hence $|\mathcal{H}_{\boldsymbol{\phi}}(x_{1:n})| \leq (n+1)^p$.

Now $\mathcal{H}_{\boldsymbol{\phi}}(x_{1:n})$ only depends on $\{(\phi_j(x_{1j}), \ldots, \phi_j(x_{nj})) : j = 1, \ldots, p\}$ and for each j, we already know that $|\{(\phi_j(x_{1j}), \ldots, \phi_j(x_{nj})) : \phi_j \in \mathcal{B}\}| = |\mathcal{B}(x_{1j}, \ldots, x_{nj})| \leq (n+1)^2$. Thus

$$|\{((\phi_j(x_{1j}), \ldots, \phi_j(x_{nj})) : j = 1, \ldots, p) : \phi_j \in \mathcal{B},\, j = 1, \ldots, p\}| \leq (n+1)^{2p}$$

whence

$$|\mathcal{H}(x_{1:n})| \leq (n+1)^{2p} \times (n+1)^p = (n+1)^{3p}.$$

Result: if $z_i = (x_i, y_i)$, then

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \leq \sqrt{\frac{2 \log |\mathcal{H}(x_{1:n})|}{n}}.$$

Taking expectations then

$$\mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{6p \log(n+1)}{n}}.$$

Thus

$$\mathbb{E} \sup_{h \in \mathcal{H}} \{R(h) - \hat{R}(h)\} \leq 2\sqrt{\frac{6p \log(n+1)}{n}}.$$